Suvivuo, Sampsa

Qualitative Big Data's Challenges and Solutions

Published in:
Proceedings of the 54th Hawaii International Conference on System Sciences

Published: 05/01/2021

Document Version
Publisher's PDF, also known as Version of record

Qualitative Big Data's Challenges and Solutions: An Organizing Review

Sampsa Suvivuo
Aalto University School of Business
sampsa.suvivuo@aalto.fi

Abstract

Digitalization of everyday lives has tremendously increased the amount of digital (trace) data of people's behaviour available for researchers. However, traditional qualitative research methods struggle with the width and breadth of the data. This paper reviewed 61 recent studies that had utilized qualitative big data for the practical challenges they had encountered and how they were addressed. While quantitative and qualitative big data share many common issues, the review points at that lack of qualitative methods and dataset reduction required by algorithms in big data research decreases the richness of the qualitative data. Locating relevant data and reducing noise are further challenges. Currently, these challenges can be only partially addressed with a combination of human and computer pattern recognition and crowdsourcing. The review describes many "tricks of the trade" but abduction research and pragmatist philosophy seem promising starting places for a more pervasive framework.

1. Introduction

The amount of data in the world doubles every two years [1]. 80% or so of companies' data is qualitative or unstructured i.e. text (emails, web pages, social media, blogs, documents), video, audio and images [2-3]. This has led to a situation where there is more data available than traditional qualitative tools can cope with [4-6]. In big data research (BDR), sample sizes might be measured in millions and a system could produce petabytes worth of data in a matter of hours. Data is often user-generated rather than explicitly created, collected and stored for research purposes. In 2018, 16% of papers in top IS journals could be classified as big data research [7].

Like any other line of inquiry, big data has its limitations. Many of which are shared by quantitative and qualitative data alike. For quantitative research, big data caused the need to address the "deflated p-value" issue. Many statistical means were developed for smaller samples and with very large samples "*the immense volume of data means that almost everything is significant*" meaning that with big data, the p-value alone is not sufficient to determine if the results are significant [8-9]. Since the issues common to quantitative and qualitative big data have been extensively described elsewhere (more on this in the next section), this review focuses on the question: "What issues researchers have encountered with qualitative big data and how these issues can be addressed?"

Rather than engaging in gap spotting, the review's purpose is to act as an organizing review describing and synthesizing challenges encountered with qualitative big data [10]. This review tries to distinguish between challenges encountered in qualitative research in general such as difficulties in generalizing results from one setting to another or the need for domain knowledge, and challenges unique to qualitative big data.

So far, the majority of ISS research has been built on small data. However, big data can uniquely advance the development of theory by revealing anomalies, alternative conceptualizations of constructs and new field experiments. [7]. Interviews and surveys are artificial situations, but discussion forums and other digital sources contain records with candour and the source is also updated and the topics evolve [36]. Large scale interview and survey studies can also be slow and prohibitively costly. Studies into asthma risk factors usually involve one or two triggers, but Zhang and Ram [37] could compare 270 risk factors simultaneously, while determining their relative importance, capturing rarely studied environmental factors. Many researchers agree [9, 13, 14] that big data and small data are not mutually exclusive but complementary. Sampling big data is only partly satisfactory. Often it is not clear beforehand which part of the dataset contains the most interesting data and a small sample size may cause the researcher to miss temporal shifts and other aggregates. To study social

URI: https://hdl.handle.net/10125/70731
978-0-9981331-4-0

patterning or collective expression of the phenomenon, limits set by manual coding must be overcome. [15].

The paper proceeds as follows. First, big and small data are defined, followed by a short description of issues common to qualitative and quantitative big data. The next section presents how the literature review was conducted regarding the selection of articles and coding. Then, identified issues and their mitigation strategies are examined followed by a discussion on the review's results, implications, limitations and further studies.

1.1. Big and small data

The nature of big data and what constitutes it is still discussed today [16], but often it is described with the four Vs of volume, velocity, variety and veracity [17-18]. The volume stands for an enormous quantity of data by the disciplines' standards. Velocity means the dataset is not static but collected in real-time or at least updated regularly. Big data is often collected from multiple sources or the dataset contains structured and unstructured data simultaneously introducing variety and richness. Finally, big data is often "noisy" requiring preparation before analysis or there could be other forms of uncertainty in the data, for example regarding how the data was collected or stored. For their study, Kitchin and McArdle [19] amended the four Vs with exhaustivity (n=all), resolution (fine-grained), indexicality (identification), relationality (common fields that allow conjoining with other data), extensionality (it is easy to add or change the fields) and scalability (expanding rapidly in size). After the study of 26 datasets from seven different domains, they determined that velocity and exhaustivity were the key attributes in determining whether the dataset is big data or not.

Closely linked to big data are digital trace data, records of activity in information systems. Whether digital or analog, the trace data are by-products of actions i.e. not especially produced for research purposes but are "found". They are event-based and since these events occur over time, they are longitudinal. These properties make them different, for example, from survey or interview data. Archival data might be trace data as long as it is not summary data. [20-21]. "*Trace data are created, not given*" and are often semi-structured [22]. For example, a post in Facebook's Timeline contains a timestamp, user-ID, possibly location data together with unstructured text and/or picture/video.

For this paper, we define big data as a dataset so large, its manipulation and analysis by manual means is not feasible. For the small data, we adopt a definition by Kitchin and McArdle as "*data that have been produced in tightly controlled ways using sampling techniques that limit their scope, temporality and size, and are quite inflexible in their administration and generation*", [19].

1.2. Common challenges with big data

Earlier research has identified many issues regarding big data. A researcher might think big data is automatically better than small data regardless of the research question due to "Big data hubris" [23]. Big Data's representativeness is called into question asking if it really represents people in general or just people with access to the internet and smartphones [21, 24, 25]. Use of a proxy is often necessary but big data can also become a convenience sample collected not because it was the best approach but because it was easier, faster and cheaper to collect [4, 5, 21, 26]. Because the way datasets are collected and presented is not uniform, combining datasets is difficult [12, 16]. With a dataset big enough almost every relationship will become statistically significant causing spurious correlations [7, 18, 27]. Big data research tends to focus on the "tactical" issues at the expense of "why" settling for correlations [7]. In general, big data is better suited to providing the "what", the "where" and the "when" but not the "why" or the "how" [14, 18].

Blindly collecting data or using a dataset collected by someone else might lead to the loss of context and circumstances the data was created in [5, 11, 16, 19, 24, 25, 28]. Most of the big data is created, collected and owned by corporations leading to the "big data divide" i.e. different access to big data between researchers and researchers operating with "data fumes" [7, 28, 29]. The research infrastructure or apparatus required to create, collect, store and analyse big data is influenced by sociotechnical aspects meaning that the big data is not as objective as initially thought [20, 22, 25, 30]. The black-box nature of APIs also threatens the replicability of studies and the reproduction of datasets [16, 23, 31]. Researchers also tend to work under the "Ideal User Assumption" assuming all the users are operating in good faith and not trying to game the system or engaging in any opportunistic behaviour [21, 23]. Big data research must also address ethical questions regarding informed consent, minimization of harm, anonymization, privacy and searchability of participants [11, 20, 32]. As shown by Daries et al. [33], anonymization causes distortions in the data. All these issues are not specific to big data studies, but because of the size of datasets in big data research, they are more pronounced and their consequences potentially more severe than with traditional "small data" studies [25].

2. Literature review

An organizing review [10] was conducted to identify the practical challenges researchers have had with qualitative big data and how the challenges were addressed. The review's steps include a manual staged review with articles from the AIS Senior Scholars' Basket of eight journals (European Journal of Information Systems, Information Systems Journal, Information Systems Research, Journal of AIS, Journal of Information Technology, Journal of MIS, Journal of Strategic Information Systems and MIS Quarterly) and six journals with explicit big data focus (Big Data & Society, Big Data and Information Analytics, Big Data Research, Frontiers in Big Data, IEEE Transactions of Big Data, Journal of Big Data), followed by a selection of articles, their review and finally a synthesis of the findings.

2.1. Search and selection of articles

Qualitative big data and its issues are a universal phenomenon but to narrow the breadth, this review focuses on the issue from the ISS point of view. The search term "big data" produced 9 552 results in the AIS's electronic library and many more in other databases that include other disciplines as well. 77 774 results in EBSCOhost (across all databases), 81 653 in Scopus and 110 352 in Web of Science (across all databases) respectively. Guided knowledge discovery in the form of keywords requires prior knowledge (or a hunch) of suitable search terms. This approach was initially attempted but terms such as "CAQDAS" (Computer Assisted Qualitative Data Analysis), "qualitative big data" and "qualitative" AND "big data" proved out to be not accurate enough resulting either in no results at all or in a very disparate collection of articles.

For the lack of search terms, a staged review of titles and abstracts was carried out by hand. In uncertain cases, the data collection section was studied to decide if the article should be included. Articles published between 2017 and 2020 in the journals belonging to the AIS's Senior Scholars' Basket of Journals and six big data journals acted as an initial dataset of 1876 publications. The review focused on journal papers as the longer format allows authors more room to give more details and minutia of their research.

To be included in the review, the article had to be empirical and to utilize qualitative big data. Text, videos, pictures and audio in themselves instead of aggregates or metadata such as number of posts, ratings, string length, retweets, likes, votes, follower count or similar measures. The review began with AIS Basket and there the biggest completely hand-coded sample the review came across was 23 000 tweets [34]. To ensure an adequate difference between big and small data studies and buffer to [34] while simultaneously not being too high to exclude big data studies from the smaller end of the scale, the threshold for observations was set at 30 000.

Since sources and types of data are not equal in how laborious or demanding their study is, an exception was made for the [44] which contains an analysis of 19 873 YouTube videos. For example, evaluating whether a discussion forum data constitutes big data by post count alone is difficult as posts' wordcount can vary from one to over a thousand words as in [35].

Finally, when working with qualitative big data, researchers are bound to turn to quantitative methods to analyse it. Thus, the emphasis was that the data was qualitative, not the methods. Of the 1876 articles published in the selected journals between 2017 and 2020, 61 were included in the review. However, due to space constraints, only the articles seen as the most versatile and illustrative [36-65] are discussed here. The complete list of articles is available upon request from the author.

2.2. Review and coding of articles

A coding scheme captured each article's keywords and phrases, data source(s), initial sample size (because finding relevant data among big data is one of its challenges), perceived issues and possible mitigation strategies. Among the 345 keywords "big data" occurred 11 times (3,2%) and "qualitative" only twice. Data sources were captured to examine the diversity of sources – 15 studies collected all of their data from Twitter and 5 used Twitter data among other sources. Sample sizes help to ascertain that the study can be described as big data research. Shorter texts numbered from hundreds of thousands to millions while longer texts such as loan applications, petitions and news articles numbered from 30 000 to 50 000. Perceived challenges and their counters are vital for the review's objective. The challenge might be based on the research question, stated in the text or be inferred. Finally, challenges and their solutions were categorized under common themes.

3. Identified challenges and their counters

It is not customary that articles contain reports on hardships with data analysis. Thus, identifying issues often required "reading between the lines" and

interpreting why certain methods were chosen and actions taken. For example, when using several keyword searches and aligning discussion forum data with other events and by utilizing both automatic text-mining tool Leximancer and manually operated NVivo, McKenna [36] was reducing the dataset to a more manageable size expressing simultaneously the issue and its counter.

As per our definition of big data, many of the issues stem from the fact that datasets are so large, they cannot be manipulated or analysed manually. Articles in this review address this either by "making big data small" by zooming on relevant data or by scaling-up their qualitative research i.e. annotated ground truth, with the help of machine learning. Though humans can account for many issues shown below almost on a subconscious level, the machine must be explicitly told how to address each issue. A summary of identified issues and their counters can be seen in table 1 within section 3.4.

3.1. Lack of qualitative tools

Large unstructured corpora must be (at least partially) transformed into structured before (quantitative) analysis. However, this tends to be labour intensive (expensive) meaning that the annotated training data are small compared to the rest of the data [37]. This is especially true with highly domain-specific medical information requiring several expert annotators [59].

Theories guiding the research might also be from "pre-online times" as observed by [36] with social movement theories. As noted in [38], there is a need for future research to "*develop new theories for capturing linguistic and other patterns in the rich, abundant content generated by ICT communication*". They also note that a further study combining quantitative and qualitative analysis is needed of the factors influencing petition success and recommend focusing on videos and pictures related to the petitions. Thus, implying that a qualitative study is needed to gain more in-depth knowledge. Chen et al. [39] also call for more field studies, experimental designs and netnographic studies.

As will be shown in section 3.3., crowdsourcing has been used with success to annotate data. However, this seems to be the only qualitative technique used in these studies in addition to researchers manually coding training sets themselves. Before new qualitative tools and theories are developed, a mixed-methods approach combining both qualitative and quantitative worldviews could either act as a stopgap solution or as a foundation on which new approaches can be built. It is noteworthy, that none of the selected studies incorporated surveys, interviews or similar measures to supplement the study.

3.2. Tool induced lack of depth

Qualitative data and research are denoted by "rich descriptions", but quantification of qualitative data reduces this richness. For example, a petition might be unreasonable to begin with and this is something the algorithms cannot account for [38]. The unsupervised tools often focus on term frequencies and might miss less frequent terms meaning that the result is not representative of the whole content but a picture of what is popular [41, 62]. This was solved in [41] by categorizing content according to its similarities and then sampling it. Supervised methods, on the other hand, are limited by the requirement of predefined number of topics, which is why interesting, hidden topics could be missed [40]. However, the number of topics can be tuned to triangulate and address this.

Sentiment analysis is limited to classifying text into positive, neutral and negative. For example, 43 550 product ideas' feedback valences were categorized into positive and negative [42]. What was lost, or at least left outside of the article, was more detailed information on the nature and common denominators of positive or negative feedback. This can be alleviated (to a degree) by using aspect-based sentiment analysis to connect the sentiment to a particular aspect [43, 63] or by using the Apriori algorithm to establish association rules between sentiments and different issues [65].

3.3. Noisy data

Although there is noise in both quantitative and qualitative big data, it can be argued that qualitative noise can be harder to address. Quantitative methods have tools to address missing values, outliers and similar issues. While strongly connected to finding relevant information, annotating can also be thought of as a form of noise reduction. Without annotated "ground truth" or "golden standard" data, many algorithms lose a lot of their utility. With big data volumes, even relevant data becomes noise unless it can be identified.

Informal language, abbreviations, misspellings, punctuation errors, non-dictionary slang, wordplay, comparative sentences, negation, transferred negation, double negation, sarcasm, unwanted languages, spam and emoticons constitute noise in texts [37, 43, 44]. Words related to the topic could be used in ways that didn't necessitate their inclusion adding noise such as*: "I will call you asthma, because you take my breath*

away" [37]. Because the language is continuously evolving, lexicon- or knowledge-based methods cannot keep up with it [44].

This noise is often addressed in pre-processing by reducing messages' features [45]. Still, uppercase and extended words (i.e. *FUNNYYY, looool*) can be used to decipher sentiment instead of just being removed as noise [61]. To spot informal language in [64], the data was verified against a dictionary of 1 million distinct words to spot the out-of-vocabulary words for further examination.

Several different conversations might be taking place within a single thread [46] or the case of "dynamic truth" where a claim is updated over time [60] can be difficult for algorithms to identify and address. When faced with multiple conversations within discussion forum threads, Abbasi et al. [46] converted 5 million posts into 26 million sentences with more focus and consistency. A similar approach was adopted in [48] when documents were analysed at a paragraph level to better identify relevant parts.

People use different words to describe the same topic, meaning that dictionaries might include multiple synonyms and polysemes which in turn increase computation requirements. Zhou et al. [47] addressed this by reducing keyword dimensions from 41 101 to 8 435 with SVD-technique making processing more efficient. A word might have different sentiment values depending on the sentence and/or context it occurs, but some approaches do not consider the order of words. [48, 63, 64]. Accuracy can be increased by joint analysis of local (word's syntactic features) and global (document, paragraph) contexts [58, 63].

Often some thousands of observations are annotated by researchers, junior faculty or workforce recruited from the student population [37, 39, 44, 45, 49, 50]. However, this might not be enough or there is a need to verify labels [44, 45]. It is also possible to crowdsource more labels for the training data based on a small initial dataset coded by the researcher or to produce a human coded dataset that the algorithm's results are compared to. However, crowdsourcing annotation has challenges regarding label noise, intercoder reliability and allocation [51, 57].

3.4. Finding relevant data

Big data is voluminous, but again, it can be argued that challenges with qualitative data are unique compared to quantitative data which is usually structured, and it is known what it's depicting. For example, for their *10 gigabytes* of *relevant plain text data*, [52] also collected "*a nontrivial amount of irrelevant data*". Another study [49] managed to narrow the 1,25 million results for one keyword to 131 759 blog posts for all keywords by using a specialized search engine. When key terms do not occur simultaneously, arranging terms under topics can help to find relevant documents [56, 62]. Although a lot of research on online health communities (OHC) has used a manual content analysis approach, the approach becomes quickly unfeasible when the volume increases. [39] manually coded 3 086 replies from OHC's forum, used them as training data for the support vector machine, and finally used a trained classifier to code the remaining replies. Still, a machine learning approach is not immune to misclassifications [43].

Liu et al. [44] wanted to study videos uploaded to YouTube to assess, on a scale, whether they contained a high or low degree of diabetes-related information. At the time of the study, YouTube had over 100 million videos. Making use of the videos' metadata, 19 873 videos of interest were identified using 200 keywords derived from the discussions on OHC. They also made use of videos' captions, but those were available for only 11% of videos in the sample. Another approach is to align the search with events of interest. A study of hacker platforms had to locate relevant content among 2 960 893 posts in 355 222 threads and found posts relevant for the study by looking for posts with mentions of the same port numbers listed by threat databases [53]. A similar approach was adopted by [36] when a search was centred around game patches.

The use of readymade dictionaries and the creation of data specific dictionaries are ways to analyse the contents and to generate keywords. To filter customer complaints from compliments and messages seeking or sharing information [50] adopted a lexicon approach with 326 complaint n-grams and 354 compliment n-grams. Still, general-purpose lexicons might perform badly in capturing nuances in domain-specific contexts. Medical dictionaries were needed for breaking unstructured questions and answers into entities [59] and mapping health-related terms in messages against professional health terminologies [39]. Noting that constructing a dictionary manually might miss colloquial and informal terms and necessitate dimension reduction or focusing on a smaller sample, the Naïve Bayes classifier was used to create a dictionary based on Yelp reviews [54]. This way, words such as "pungency" and "wiping nose" that authors would not have realized to include in the dictionary, were included.

Table 1. Identified issues and their counters by theme

	Challenge	Counter
Lack of qualitative tools	There is no viable way to analyse a large qualitative corpus without turning it into quantitative data first.	Use crowdsourcing [45, 50] and mixed-method approach as a stopgap solution before the development of new theories and techniques.
Tool induced lack of depth	When content is classified into three sentiments (for example) some of the “rich description” is inevitably lost.	Looking beyond the usual tools [38], generating dictionaries from the data to supplement readymade dictionaries [54], using aspect-based sentiment analysis [43,63] or establishing association rules between sentiments and issues [65].
	Extracting a representative sample, not what is popular.	Categorizing content based on their similarities between each other and then taking a sample [41].
Noisy data	Informal language, abbreviations, misspellings, punctuation errors, non-dictionary slang, wordplay, emoticons, URLs.	Natural language processing, machine learning and domain adaptation [37], data pre-processing [45]. Use a dictionary to check for out-of-vocabulary words [64]
	There could be several discussions within one thread.	Break posts into sentences [46], break the document into segments to analyse at paragraph level [48].
	Consumers may have different search intent and the keywords they use in their search activities may reflect this intent.	Focus on the search goal and use a dictionary relevant to it [40].
	Issues relating to <phenomenon> may be presented in various forms as it is highly likely that people will use different terms when referring to the same topic.	Construct the dictionary directly from the text with Naïve Bayes classifiers and apply SVD [47].
	Depending on the context, the words’ sentiment may change.	Combine sentence-level features (SLF) with domain sensitive features (DSF) [63].
Finding relevant data	Manual coding and analysis of data is unfeasible.	Manually code a small sample and use it to train a classifier [39] or verify the annotated sample by crowdsourcing [45]. Use a readymade [50] or custom dictionary [39, 44, 54, 55] and apply it to the remaining data.
	Relevant search terms do not occur simultaneously in documents.	Arrange terms under topics and use the topics to find relevant documents [56, 59, 62].
	Identifying posts of interest among millions of posts on message boards.	Connecting data to other events or sources [36,53].

A similar approach was adopted in a study on how information accumulated while helping others affects the quality of solutions by creating an information network where individuals were connected based on 115 topics generated from 2 million messages [55]. However, unless the dictionary contains more than single words (phrases up to two or three words), there is a risk of oversimplifying the language [54].

4. Discussion

In-line with the general characteristics of big data, challenges with qualitative big data stem from volume, variety and veracity. Velocity was not an issue as studies were mostly backward-looking, collecting data at certain points rather than continuously. Finding relevant data and addressing noise can be viewed as two sides of the same coin. Reducing noise helps to find relevant data and vice versa. However, the biggest challenge working with qualitative big data is likely preserving the richness of data to enable answering the "why" and the "how" questions.

Due to the nature and aim of this review, the challenges and their counters presented are mostly very practical in nature. Still, these challenges and how researchers choose to address them carry wider implications for BDR as a whole. What the researcher keeps as noise and how it is removed is not trivial and neither is the influence of context. The use of aspect-based sentiment analysis instead of regular sentiment analysis might preserve more of the data and enable richer analysis.

Certain methods such as LDA, LSA and SVM repeatedly show up in the articles. When many studies use a few selected methods, BDR might risk pipe vision and research that can only answer certain kinds of questions and deliver certain kinds of insights. A mixed-methods approach could be supplemented with abductive research iterating between discovery and justification and pragmatist philosophy as suggested by Lindberg [6] to help maintain the richness of qualitative data. Abduction would also allow BDR-researcher to address the "why" behind patterns [7]. Abbasi et al. noted that pragmatism (not what people say, but what they do with language) and language-action perspective could advance sensemaking in the social media context [66].

In their paper, Grover et al. [7] raised the issues of fishing for interesting relationships (r-hacking) and creating hypotheses after the results are known (HARKing) as threatening generalizability and value of results. However, in BDR, the first step might well be looking at the big picture and then focusing on interesting relationships. As portrayed in [6], large-scale patterns and structures are typically analysed with machine pattern recognition whereas human pattern recognition is used when zooming in human dynamics. It is noted in [7] that algorithms can be used to uncover novel patterns in data to offer *initial structural frames* for deeper theory-building via the study of small samples. Many of the reviewed studies combine human and machine pattern recognition but do not reflect on what the findings mean for the theory nor use theories to guide the annotation process and creation of "ground truth". Overall, as expressed in [7], the reviewed articles focus on "tactical" issues and do not have strong theoretical underpinnings.

5. Conclusion

The absence of qualitative methods is noticeable in qualitative big data research. Due to volume and lack of qualitative tools, unstructured data must be transformed into structured, reducing the data and rich descriptions qualitative data is known for. This also decreases the ability to answer "why" and "how" questions. There are several sources and forms of noise and finding relevant data can be difficult.

The usual approach to working with qualitative big data combines both human and machine pattern recognition in the form of a hand-coded sample, which is often verified and/or expanded by a crowdsourced workforce and used to train a classifier to label remaining data. Topic modelling, tailored dictionaries and connecting data to other events help to zoom-in on interesting phenomena. In addition to these broad lines, this review describes many "tricks of the trade" researchers have used to address specific challenges.

The amount of data in the form of texts, videos, audio and pictures is likely to increase in the future [67] and the research community must answer this development. Not necessarily by adapting existing tools and theories but by creating new ones. Studies included in this review combine economics, machine learning, statistical methods, linguistics and many other fields using method after method and creating hybrids to address challenges, so it is to be expected that the new methods to be as complex as well. In the future, the division between qualitative and quantitative research might become increasingly tenuous as posited in [6]. A more sensible division – if one must be made – might be distinguishing between small and big data parts for the sake of iteration and abduction.

6. Further studies and limitations

After examination of keywords, it becomes apparent that instead of the type of terms initially tried, the proper search terms for systematic qualitative big data review would be sources associated with qualitative big data (Twitter, Facebook, Youtube, Amazon, Wikipedia, discussion boards), methods used to analyse it (text- or opinion-mining, word embedding, topic modelling, crowdsourcing) or topics related to consumers and users (e-commerce, crowdsourcing, reviews, MOOCs) as opposed to business or professional organizations.

The review is (knowingly) bounded within the ISS and big data journals. Still, due to the universal nature of qualitative big data and its challenges, the review should be expanded to other disciplines such as digital humanities. Interviewing or surveying researchers could also yield practical challenges not caught by the review such as possible lack of necessary skills or network to engage in big data research or not seeing the value in it.

We acknowledge several limitations in this paper. Despite the high initial number of articles, the final sample was very small. A staged review focusing on headlines and abstracts has likely missed articles that should have been included or the 30 000-observation threshold could have been too high. Many articles were excluded because the amount of data they used could not be ascertained. Most of the studies in the review use texts as their source while video or images are used by some, but audio is used by none. This might simply represent the current state of qualitative big data usage or it might be a result of sampling.

10. References

[1] IDC (2014). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things.https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. Accessed 17.2.2020.

[2] Kuehcler W., L. 2007. Business Applications of Unstructured. Communications of the ACM October 2007/Vol. 50, No. 10.

[3] P. Russom, "BI Search and Text Analytics: New Addition to the BI Technology Stack," TDWI Best Practices Report, The Data Warehousing Institute, Renton, WA, USA, Tech. Rep., Second Quarter, 2007.

[4] Berente, N., and Seidel, S. (2014). Big data and inductive theory development: Towards computational Grounded Theory? 20th Americas Conference on Information Systems, AMCIS 2014, 1–11.

[5] Walsh, I., Holton, J. A., Bailyn, L., Fernandez, W., Levina, N., and Glaser, B. (2015). What Grounded Theory Is…A Critically Reflective Conversation Among Scholars. Organizational Research Methods, 18(4), 581–599. https://doi.org/10.1177/1094428114565028

[6] Lindberg, A. (2020). Developing theory through integrating human and machine pattern recognition. Journal of the Association for Information Systems, 21(1), 90–116. https://doi.org/10.17705/1jais.00593

[7] Grover, V., Lindberg, A., Benbasat, I., and Lyytinen, K. (2020). The Perils and Promises of Big Data Research in Information Systems. Journal of the Association for Information Systems, 21, 1–26. https://doi.org/10.17705/1jais.00601

[8] Lin, M., Lucas, H. C., and Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. Information Systems Research, 24(4), 906–917. https://doi.org/10.1287/isre.2013.0480

[9] George, G., Haas, M., R., Pentland, A., (2014). From the Editors. Big data and management. Management Journal 2014, Vol. 57, No. 2, 321–326. http://dx.doi.org/10.5465/amj.2014.4002

[10] Leidner, D. E. (2018). Review and theory symbiosis: An introspective retrospective. Journal of the Association for Information Systems, 19(6), 552–567. https://doi.org/10.17705/1jais.00501

[11] Chang, R. M., Kauffman, R. J., and Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. Decision Support Systems,63,67–80. https://doi.org/10.1016/j.dss.2013.08.008

[12] Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. BioScience, 59(7), 613–620. https://doi.org/10.1525/bio.2009.59.7.12

[13] Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data and Society, 1(1). https://doi.org/10.1177/2053951714528481

[14] Dalton, C., and Thatcher, J. (2014). What Does a Critical Data Studies Look Like, and Why Do We Care? Psychological Bulletin, Vol. 126, p. 21. https://doi.org/.1037//0033-2909.I26.1.78

[15] Karamshuk, D., Shaw, F., Brownlie, J., and Sastry, N. (2017). Bridging big data and qualitative methods in the social sciences: A case study of Twitter responses to high profile deaths by suicide. Online Social Networks and Media,1,33–43. https://doi.org/10.1016/j.osnem.2017.01.002

[16] Jones, M. (2019). What we talk about when we talk about (big) data. Journal of Strategic Information Systems, 28(1), 3–16. https://doi.org/10.1016/j.jsis.2018.10.005

[17] Goes, B., P., (2014). Editor's Comments – Big Data and IS Research. MIS Quarterly Vol. 38 No. 3 pp. iii-viii/September 2014

[18] Abbasi, A., Sarker, S., and Chiang, R. H. L. (2016). Big data research in information systems: Toward an inclusive research agenda. Journal of the Association for Information Systems, 17(2), 1–32. https://doi.org/10.17705/1jais.00423

[19] Kitchin, R., and McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics

of 26 datasets. Big Data and Society, 3(1). https://doi.org/10.1177/2053951716631130
[20] Howison, J., Wiggins, A., and Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. Journal of the Association for Information Systems, 12(12), 767–797. https://doi.org/10.17705/1jais.00282
[21] Lazer, D., and Radford, J. (2017). Data ex Machina: Introduction to Big Data. Annual Review of Sociology, 43(1), 19–39. https://doi.org/10.1146/annurev-soc-060116-053457
[22] Østerlund, C., Crowston, K., & Jackson, C. (2020). Building an apparatus: Refractive, reflective, and diffractive readings of trace data. Journal of the Association for Information Systems, 21(1), 1–22. https://doi.org/10.17705/1jais.00590
[23] Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. Science, 343(6176), 1203–1205. https://doi.org/10.1126/science.1248506
[24] boyd, danah, and Crawford, K. (2012). Six Provocations for Big Data. SSRN Electronic Journal, 1–17. https://doi.org/10.2139/ssrn.1926431
[25] Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? Qualitative Research, 18(6),591–603. https://doi.org/10.1177/1468794117743465
[26] Agarwal, R., and Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. Information Systems Research, 25(3), 443–448. https://doi.org/10.1287/isre.2014.0546
[27] Strong, C. (2014). The challenge of "Big Data": What does it mean for the qualitative research industry? Qualitative Market Research, 17(4), 336–342. https://doi.org/10.1108/QMR-10-2013-0076
[28] Thatcher J (2014) Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. International Journal of Communication 8 (2014), 1765–1783.
[29] Dalton, C. M., Taylor, L., and Thatcher, J. (2016). Critical Data Studies: A dialog on data and space. Big Data and Society, Big Data and Society January–June 2016: 1–9. https://doi.org/10.1177/2053951716648346
[30] Crawford, K., Gray, M., and Miltner, K. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. International Journal of Communication, 8, 1663–1672. Retrieved from http://ijoc.org/index.php/ijoc/article/download/2167/1164
[31] Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. Big Data and Society, 3(1), 1–15. https://doi.org/10.1177/2053951716645828
[32] Webb, H., Jirotka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., Burnap, P. (2017). The ethical challenges of publishing Twitter data for research dissemination. WebSci 2017 - Proceedings of the 2017 ACM Web Science Conference, 339–348. https://doi.org/10.1145/3091478.3091489
[33] Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., … Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. Communications of the ACM, 57(9), 56–63. https://doi.org/10.1145/2643132
[34] Vaast, E., Safadi, H., Lapointe, L., and Negoita, B. (2017). Social media affordances for connective action: An examination of microblogging use during the Gulf of Mexico oil spill. MIS Quarterly: 41(4), 1179–1206. https://doi.org/10.25300/misq/2017/41.4.08
[35] Guo, K. H., and Yu, X. (2020). The anonymous online self: Toward an understanding of the tension between discipline and online anonymity. Information Systems Journal, 30(1), 48–69. https://doi.org/10.1111/isj.12242
[36] Mckenna, B. (2019). Creating convivial affordances: A study of virtual world social movements. Information Systems Journal. 2020; 30:185–214. https://doi.org/10.1111/isj.12256
[37] Zhang, W., and Ram, S. (2020). A Comprehensive Analysis of Triggers and Risk. MIS Quarterly Vol. 44 No. 1, pp. 305-349/March 2020. https://doi.org/10.25300/MISQ/2020/15106
[38] Chen, Y., Deng, S., Kwak, D., Elnoshokaty, A., and Wu, J. (2019). A Multi-Appeal Model of Persuasion for Online Petition Success: A Linguistic Cue-Based Approach. Journal of the Association for Information Systems (2019) 20(2), 105-131. https://doi.org/10.17705/1jais.00530
[39] Chen, L., Baird, A., and Straub, D. (2019). Fostering Participant Health Knowledge and Attitudes: An Econometric Study of a Chronic Disease-Focused Online Health Community. Journal of Management Information Systems, 36:1, 194-229 https://doi.org/10.1080/07421222.2018.1550547
[40] Gong, J., Abhishek, V., and Li, B. (2018). Examining the Impact of Keyword Ambiguity on Search Advertising Performance. MIS Quarterly Vol. 42 No. 3, pp. 805-829/September 2018. https://doi.org/10.25300/MISQ/2018/14042
[41] Guo, X., Wei, Q., and Chen, G. (2017). Extracting Representative Information on Intra-Organizational Blogging Platforms. MIS Quarterly Vol. 41 No. 4, pp. 1105-1127/December 2017.
[42] Liu, Q., Du, Q., Hong, Y., Fan, W., & Wu, S. (2020). User idea implementation in open innovation communities: Evidence from a new product development crowdsourcing community. Information Systems Journal, (April 2018), 1–29. https://doi.org/10.1111/isj.12286
[43] Ho, S. Y., Wai, K., Choi, S., and Yang, F. F. (2019). Harnessing Aspect-Based Sentiment Analysis: How Are Tweets Associated with Forecast Accuracy? Journal of the Association for Information Systems (2019) 20(8), 1174-1209 https://doi.org/10.17705/1jais.00564
[44] Liu, X., and Zhang, B. (2020). Go to Youtube and Call Me in The Morning. MIS Quarterly Vol. 44 No. 1, pp. 257-283/March 2020. https://doi.org/10.25300/MISQ/2020/15107
[45] Li, T., Dalen, J. Van, and Rees, P. J. Van. (2018). The information content of stock microblogs on financial markets. Journal of Information Technology (2018) 33, 50–69. https://doi.org/10.1057/s41265-016-0034-2
[46] Abbasi, A., Li, J., Adjeroh, D., Abate, M., and Zheng, W. (2019). Don't Mention It? Analyzing User-Generated Content Signals for Early Adverse Event Warnings.

Systems Research 30(3):1007-1028. https://doi.org/10.1287/isre.2019.0847
[47] Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W., and Yan, X. (2018). Measuring Customer Agility from Online Reviews Using Big Data Text Analytics. MIS Quarterly Vol. 44 No. 1, pp. 305-349/March 2020. https://doi.org/10.1080/07421222.2018.1451956
[48] Zhang, J.-B., Sun, Y.-X., & Zhan, D.-C. (2017). Multiple-instance learning for text categorization based on semantic representation. Big Data & Information Analytics, 2(1), 69–75. https://doi.org/10.3934/bdia.2017009
[49] Luo, X., and Gu, B. (2017). Expert Blogs and Consumer Perceptions of Competing Brands. MIS Quarterly Vol. 41 No. 2, pp. 371-395/June 2017
[50] Gunarathne P., Rui, H., and Seidmann, A. (2018). When Social Media Delivers Customer Service: Differential Customer Treatment in the Airline Industry. MIS Quarterly Vol. 42 No. 2, pp. 489-520/June 2018. https://doi.org/10.25300/MISQ/2018/14290
[51] Wang, Z., Jiang, C., Zhao, H., Ding, Y. (2020). Mining Semantic Soft Factors for Credit Risk Evaluation in Peer-to-Peer Lending. Journal of Management Information Systems, 37:1, 282-308, https://doi.org/10.1080/07421222.2019.1705513
[52] Huang, J., Boh, W. F., and Goh, K. H. (2018). A Temporal Study of the Effects of Online Opinions: Information Sources Matter. Journal of Management Information Systems, 34:4, 1169-1202. https://doi.org/10.1080/07421222.2017.1394079
[53] Yue, W. T., and Wang, Q. (2019). See No Evil, Hear No Evil? Dissecting the Impact of Online Hacker Forums. MIS Quarterly Vol. 43 No. 1, pp. 73-95/March 2019. https://doi.org/10.25300/MISQ/2019/13042
[54] Mejia, J., Mankad, S., and Gopal, A. (2019). A for Effort? Using the Crowd to Identify Moral Hazard in New York City Restaurant Hygiene Inspections. Information Systems Research 30(4):1363-1386 https://doi.org/10.1287/isre.2019.0866
[55] Hwang, E. H., Singh, P. V., and Argote, L. (2019). Jack of All, Master of Some: Information Network and Innovation in Crowdsourcing Communities. Information Systems Research 30(2):389-410. https://doi.org/10.1287/isre.2018.0804.
[56] Dowling, M., Wycoff, N., Mayer, B., Wenskovitch, J., Leman, S., House, L., … Hauck, P. (2019). Interactive Visual Analytics for Sensemaking with Big Text. Big Data Research,16,49–58. https://doi.org/10.1016/j.bdr.2019.04.003
[57] Chen, Y., Song, Q., Liu, X., Sastry, P. S., Hu, X., & Chen, Y. (2020). On Robustness of Neural Architecture Search Under Label Noise. 3 (February), 1–9. https://doi.org/10.3389/fdata.2020.00002
[58] Meng, Y., Huang, J., Wang, G., Wang, Z., Zhang, C., & Han, J. (2020). Unsupervised Word Embedding Learning by Incorporating Local and Global Contexts. Frontiers in Big Data, 3 (March), 1–12. https://doi.org/10.3389/fdata.2020.00009
[59] Li, Y., Liu, C., Du, N., Fan, W., Li, Q., Gao, J., … Wu, H. (2016). Extracting Medical Knowledge from Crowdsourced Question Answering Website. IEEE Transactions on Big Data, 6(2), 309–321. https://doi.org/10.1109/tbdata.2016.2612236
[60] Zhang, D., Wang, D., Vance, N., Zhang, Y., & Mike, S. (2018). On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications. IEEE Transactions on Big Data, 5(2), 195–208. https://doi.org/10.1109/tbdata.2018.2824812
[61] El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A. (2018). A novel adaptable approach for sentiment analysis on big social data. Journal of Big Data, 5(1). https://doi.org/10.1186/s40537-018-0120-0
[62] Nguyen, T., Ngoc, T., Nguyen, H., Thu, T., & Nguyen, V. A. (2019). Mining aspects of customer's review on the social network. Journal of Big Data, 1–21. https://doi.org/10.1186/s40537-019-0184-5
[63] Rintyarna, B. S., Sarno, R., & Fatichah, C. (2019). Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks. Journal of Big Data. https://doi.org/10.1186/s40537-019-0246-8
[64] Cury, R. M. (2019). Oscillation of tweet sentiments in the election of João Doria Jr . for Mayor. Journal of Big Data, 1–15. https://doi.org/10.1186/s40537-019-0208-1
[65] Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 1–16. https://doi.org/10.1186/s40537-019-0224-1
[66] Abbasi, A. (2018). Research Article Text Analytics to Support Sense-Making in Social Media: A Language-Action Perspective. MIS Quarterly Vol. 42 No. 2, pp. 427-464/June 2018. https://doi.org/10.25300/MISQ/2018/13239
[67] Shi, D., Guan, J., Zurada, J., and Manikas, A. (2018). A Data-Mining Approach to Identification of Risk Factors in Safety Management Systems. Journal of Management Information Systems, 34:4, 1054-1081, DOI: 10.1080/07421222.2017.1394056